

情感分析模型综述*

董喜双, 邹启波, 关毅*

哈尔滨工业大学, 黑龙江 哈尔滨 150001

摘要: 对情感分析模型的研究现状与进展进行了总结, 包括三个部分: 首先介绍情感分析相关概念, 包括情感分析形式化定义以及情感极性和情感强度; 然后将情感分析归纳为 3 个主要层级模型, 即词级情感分析模型、句级情感分析模型和篇章级情感分析模型。针对不同层级, 重点介绍基于无监督、有监督和半监督学习方法的情感分析模型, 并对模型的优缺点进行了详细总结; 最后介绍了情感分析模型在不同领域的应用, 包括问答系统、产品推荐和舆情分析等。

关键词: 情感分析; 情感极性; 情感强度; 机器学习

A Survey on Sentiment Analysis Models

Xishuang Dong, Qibo Zou, Yi Guan

Harbin Institute of Technology, Harbin, China, 150001

Email: dongxishuang@gmail.com; zouqibo2009@163.com; guanyi@hit.edu.cn

Abstract: This paper presents an overview of sentiment analysis models. It includes three parts: related concepts including the definition of sentiment analysis and introduction on sentiment polarity and strength; a summary of Word-level Sentiment Analysis Models, Sentence-level Sentiment Analysis Models, and Document-level Sentiment Analysis Models, which contain unsupervised based models, supervised based models, semi-supervised based models and their advantages and disadvantages; model applications such as the question and answer system, the production recommendation system, and the public opinion analysis system, respectively.

Key words: Sentiment Analysis; Sentiment Polarity; Sentiment Strength; Machine Learning

引言

心理学是研究人的心理活动及其发生、发展规律的科学^[1]。人的心理活动包括心理过程和个性, 其中心理过程分为认识过程、情感过程和意志过程三个方面。认识过程是指人的心理由表象到本质地反映客观事物心理活动; 情感过程是指人与客观事物相互作用而产生的心理活动, 包括喜、怒、哀、乐、憎、惧等情感; 意志过程是指人在目标驱使下, 克服困难并力求实现目标的心理活动。人们主要通过观察法、实验法、调查法和测验法研究人的情感过程, 但只是定性的分析人的情感过程, 而没有定量的对人的情感过程进行度量。本文总结了计算语言学领域中情感分析模型的相关研究, 详细分析了从定量的角度度量文本中情感的模型优缺点。

计算语言学中, 情感分析指判定文本所持有情感(或观点、态度)的极性和强度的过程^[2]。根据文本粒度不同, 情感性分析主要分为三个方面的研究内容: 词级情感分析 (Word-level Sentiment Analysis, WSA)、句级情感分析 (Sentence-level Sentiment Analysis, SSA) 和篇章级情感分析 (Document-level Sentiment Analysis, DSA)。词级情感分析主要包括构建情感词典、识别候选词和判断候选词情感极性与强度^[3]。情感词典构建首先人工收集种子情感词, 然后采用机器学习方法扩充种子情感词构建情感词典。在识别候选词时, 采用统计与规则相结合方法实现, 例如利用词汇共现关系识别候选词, 并且根据连词规则进一步精确识别候选词。在识别候选词的基础上, 采用无监督、有监督和半监督方式判定极性和强度。句级情感分析指判断句子的情感极性和强度^{[3][4]}。其相关过程包括: 识别情感

*基金项目: 国家自然科学基金“面向语句间相似度计算基于词主体自治学习的强化学习机制研究”, 项目编号: 60975077; 国家自然科学基金“非常规突发事件网络舆情分析方法和预警机制研究”, 项目编号 90924015。

词、分析句子结构以及判别情感极性和强度。基于无监督的句级情感分析根据句子中情感词的极性和强度累加值确定情感极性与强度。基于有监督的情感句分类是主流模型,采用的分类模型包括支持向量机(Support Vector Machine, SVM)、最大熵(Maximum Entropy, ME)以及条件随机场(Conditional Random Fields, CRFs)等;篇章级情感分析指识别篇章对某一事物的情感极性和强度^[2]。其相关过程包括:情感句分析、篇章级情感特征提取以及构建篇章级情感分类器。基于无监督的篇章级情感分析模型主要根据情感词典识别情感词,然后通过累加情感词的情感极性和强度判别篇章情感极性和强度;基于有监督的篇章情感分析主要通过提取情感特征,进而构建情感分类器来预测篇章情感极性。

情感分析不管在科学研究还是在商业应用都具有重要价值。情感分析涉及计算语言学、数据挖掘以及机器学习等方面的基础研究,处在不同学科的交叉点,因而研究情感分析可以促进不同学科的研究进展,具有重要的科学研究价值;其次,采用情感分析技术从海量数据中提取情感信息,将这些信息应用在不同领域提高服务质量,例如:网络信息安全中的不良信息过滤、电子商务中的产品推荐、搜索引擎中的个性化检索以及社交网络中的用户兴趣发掘等。

本文的组织结构如下:第二部分介绍情感分析的形式化定义以及相关要素;第三部分总结词级情感分析、句级情感分析和篇章级情感分析的相关模型及其优缺点;第四部分从问答系统、产品推荐以及舆情分析阐述情感分析技术在不同领域的应用;最后总结现有模型研究并对未来研究给予展望。为了更好的理解情感分析模型,首先介绍情感分析涉及的相关概念。

1 相关概念

给定文本 D (词汇、句子或篇章) 和情感类别集合 $C = \{c_1, c_2, c_3 \dots c_n\}$, 情感分析是将文本映射到已有的情感类别中, 该映射是一一映射, 如式 (1) 所示:

$$f: D \rightarrow C \quad (1)$$

其中, 情感类别也称情感标签 (Sentiment Labels), n 表示情感类别数, f 是情感极性判定函数, 它取决于情感分析采用的具体方法。

情感分析涉及两个重要元素: 情感极性和情感强度。情感极性是指文本对应的情感类别, 情感强度是对文本表达情感强弱的定量描述。文本情感极性通常划分为褒义 (Positive)、贬义 (Negative) 和中性 (Neutral)。此外对于具体应用, 单纯的褒贬极性分类无法满足需求, 还需要区分文本的情感强弱程度, 例如: 电子商务中的产品推荐需要根据评价为产品划定等级, 这就需要了解情感极性以及定量分析情感强度。在了解情感分析的定义及相关要素后, 下面详细介绍情感分析相关模型。

2 情感分析模型

2.1 词级情感分析模型

词级情感分析是句级情感分析和篇章级情感分析基础。基于无监督的词级情感分析主要是通过构建种子词集合, 然后通过计算种子词和候选词之间的语义相似性判断情感极性和强度。基于有监督的词级情感分析首先提取候选词的情感特征, 包括同义词、反义词、对义词和词性等, 然后训练情感词分类模型, 最后预测候选词的极性和强度。基于半监督的词级情感分析借助已有标注语料, 并不断从预测集合中抽取置信度高的样本扩充训练集合, 提高情感词预测性能。

2.1.1 基于无监督的词级情感分析模型

宋晓雷、王素格和李红霞等^[5]提出两种基于概率潜在语义分析的情感词分析模型。模型一: 首先借助概率潜在语义分析构建候选词和种子词之间的相似度矩阵, 然后利用投票法决定其情感极性, 即若候选词与种子词中积极词汇相似的数量多, 则该词是积极极性, 反之亦然; 模型二: 利用概率潜在语义分析获取候选词的语义聚类, 然后借

鉴基于同义词的词汇情感倾向判别方法判断候选词的情感倾向。该模型不借助情感语义资源实现词汇极性判断,但是数据稀疏影响相似度度量,进而影响词汇情感极性判断。Y. Wu 和 M. We^[6]将研究焦点对准上下文相关的情感形容词极性判断问题,提出将该问题转换为其修饰的名词情感极性分析,并使用基于模式和基于字符的方法来推断该名词的情感极性。基于模式的方法通过构造模板并借助搜索引擎分析词汇之间的共现数确定名词的情感极性;基于字符的方法是在语料上通过计算名词和情感词的点互信息(Pointwise Mutual Information, PMI)判断名词的情感极性,若名词与积极词汇的 PMI 值大于与消极词汇的 PMI 值,则为积极名词。此方法考虑到了形容词因为修饰对象不同而具有不同的情感倾向,但是需要编写规则模板以及借助搜索引擎来判断词的情感倾向。A. Hassan 和 D. Radev^[7]提出了基于马尔科夫随机游走模型的情感词分析方法。首先构建种子词集合,通过语义相似构建候选词和种子词网络,然后通过马尔科夫随机游走模型判断候选词的情感极性。该方法融入了词汇之间的语义关系,但是需要构建种子词的集合,并且依赖语义资源(如 WordNet)构建候选词和种子词网络。

2.1.2 基于有监督的词级情感分析模型

基于有监督的情感词分析模型流程如图 1 所示:

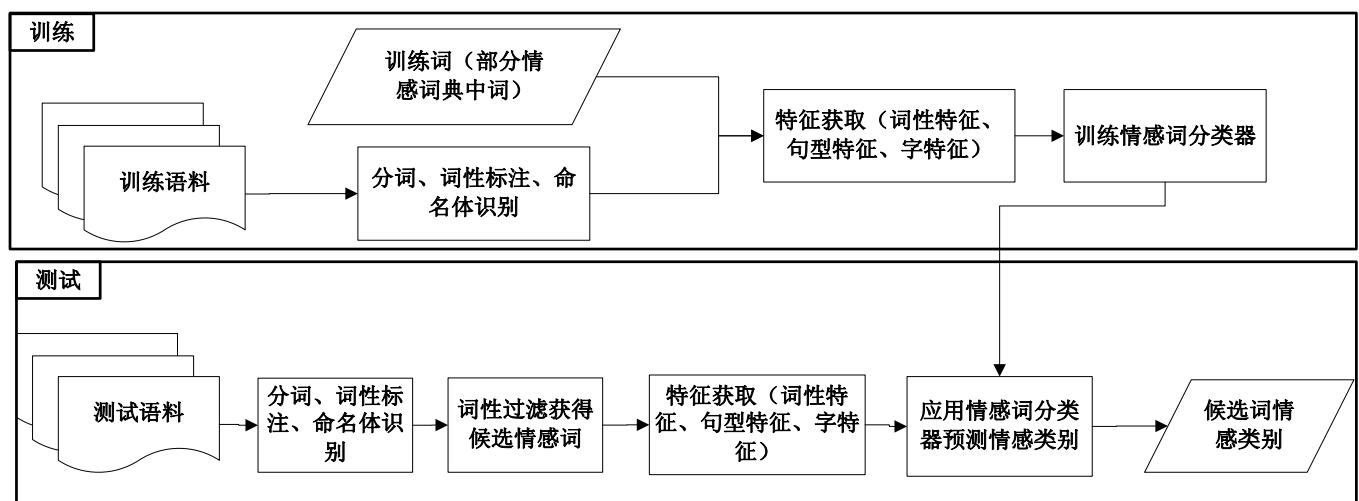


图 1 识别情感词流程图

首先对训练语料进行预处理(分词、词性标注和命名体识别),提取情感特征,训练情感词分类器,然后预测候选词情感极性。董喜双、邹启波和关毅等^[8]使用 HowNet 扩展同义词作为特征,并通过最大熵模型来预测候选词的情感极性,然后采用最小割模型来优化极性判定结果。该方法借助词汇之间的相互关系来优化结果,但由于最小割解的不唯一性,难以确定最优解。D. Das 和 S. Bandyopadhyay^[9]采用 CRFs 模型预测词汇情感极性,采用的特征包括:词性、情感词、叠字、术语或外文、特殊标点符号和句型特征(否定句、情感句和问句)等,但训练时间长。

2.1.3 基于半监督的词级情感分析模型

D. Rao 和 D. Ravichandran^[10]对比了三种半监督词级情感分析模型,分别是基于最小割的词级情感分析模型、基于随机化最小割的词级情感分析模型^[11]、以及基于标签传播(Label Propagation)的词级情感分析模型^[12],进而提出了基于词汇之间的同义关系和上下位关系的标签传播模型,完善了原有的词汇网络,提高了词汇情感极性预测精度。作者在法语和北印度语语料上验证该方法的有效性。

2.2 句级情感分析模型

句级情感分析是处于情感分析中间层,可以作为篇章级情感分析的基础。基于无监督的句级情感分析主要通过词汇之间的极性和强度累加确定句子情感极性和强度。基于有监督的句级情感分析首先提取候选句子的情感特征,

包括情感词、词序列和句型等，然后训练情感句分类模型，最后预测候选句的极性和强度。对于半监督的句级情感分析，我们重点介绍基于短语树分析模型。

2.2.1 基于无监督的句级情感分析模型

G. Fu 和 X. Wang^[13]提出基于模糊集的情感句分析模型。首先根据从细到粗的策略（Fine-to-coarse Strategy）估计句子的情感强度，具体步骤：首先构建情感词典，词典中每一个词都赋予情感强度值，然后获取句子的情感词和情感短语，最后累加句子情感词和情感短语的情感强度计算句子情感强度，以及判定情感极性；然后基于模糊分布构建三个模糊集隶属函数，分别度量句子情感极性（积极、消极和中性）的隶属程度，最后通过最大隶属原则来判断句子的情感极性。该方法实现对情感强度的模糊性进行建模，但情感词可能会因为语境不同而呈现不同的极性，无法准确判定。A. Meena, T. Prabhakar 和 G. Amati 等^[14]重点考虑了连词对句子情感极性分析的影响，结合短语和连词分析句子情感极性。但系统依赖人工构建情感词典，并且需要人工构建连词规则，不具有领域适应能力。

2.2.2 基于有监督的句级情感分析模型

基于有监督的情感词分析流程如图 2 所示：

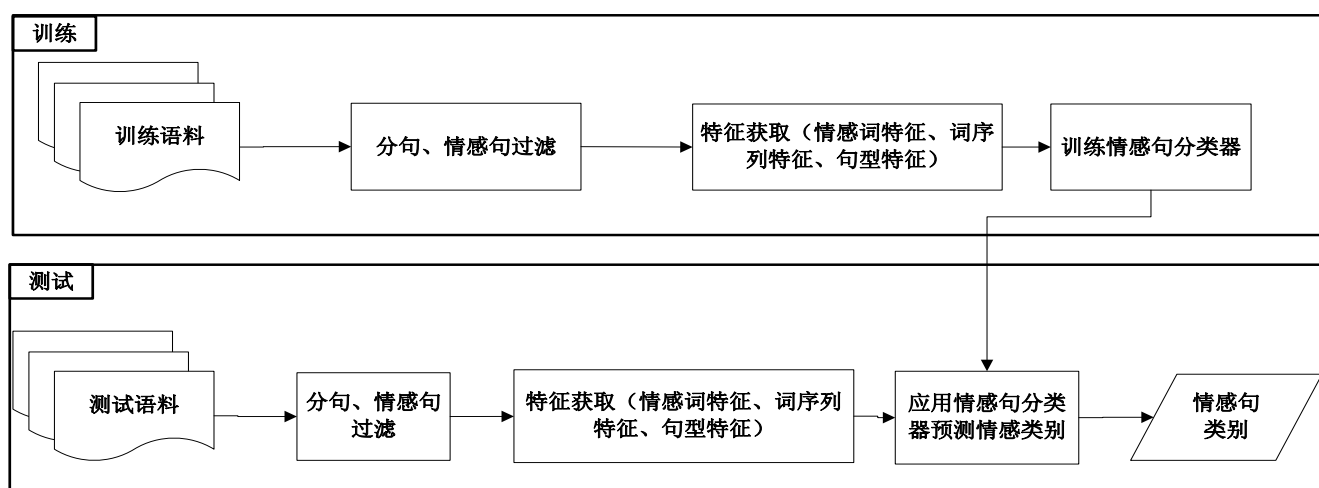


图 2 情感句分析流程图

首先构建情感句分类器，对训练语料进行预处理（分词、词性标注、命名体识别以及分句等），进而提取情感特征，训练情感分类器，然后预测句子情感极性。H. Guo, H. Zhu, Z. Guo 和 Z. Su^[15]采用无监督方法从无标注的语料中抽取各种领域专属的多级潜在情感线索，进而提出句子情感分析的领域适应模型。该模型的基本流程：（1）构建源领域和目标领域的产品特征之间的潜在关系；（2）构建目标领域情感词典；（3）利用目标领域情感词典获取目标领域多级情感线索；（4）使用多级情感线索调节源领域分类器适应目标领域，即根据目标领域多级情感线索重新训练分类器。多级潜在情感线索包括产品特征的潜在关系、领域专属和通用的情感线索。具体包括：高频词特征和一般的情感特征（情感词、情感词数量、否定词或否定标记、领域专属情感词、领域专属情感词数量和情感词分布）。R. Xia 和 C. Zong^[16]提出抽取词汇的共现关系特征构建句子情感分类器。作者提出两种快速获取词汇共现关系特征的方法：快速互信息 (Fast Mutual Information, FMI) 和快速信息增益 (Fast Information Gain, FIG)。假设词汇共现关系之间相互独立，那么两个共现词 w 和 s 与类别 c 的互信息 $I(ws, c) \approx I(w, c) + I(s, c)$ ，同理这两个词的信息增益 $GS(ws) \approx GS(w) + GS(s)$ 。该方法降低了计算复杂度，并且通过降维处理数据稀疏问题。B. Wei 和 C. Pal^[17]将研究焦点对准跨领域情感句分析，并借助机器翻译技术和结构对应学习 (Structural Correspondence Learning, SCL) 模型处理情感句的领域适应问题；为了减小机器翻译引入的噪声信息，只使用翻译结果的关键部分并且使用结构对应学习方法来寻找两种语言共享信息，但是依赖于机器翻译技术。D. Davidov, O. Tsur 和 A. Rappoport^[18]采用 K-

最邻近法 (k-Nearest Neighbors, kNN) 分析短文本 (句子) 的情感极性。情感特征分别是: 词特征、n 元语言模型特征、标点符号特征和搭配特征, 但短文版情感特征稀疏严重。W. Wei 和 J. A. Gulla^[19]将研究焦点对准产品评价句情感分析问题, 提出了基于情感本体树的层次学习方法。首先标出商品的属性以及属性具有的情感倾向, 然后将句的情感分析转换成分层的分类问题, 最后采用基于情感本体树的分层算法判断情感极性。该模型有效利用了产品属性的分层信息以及评价句中产品属性和情感倾向关系。H. Jang 和 H. Shin^[20]将研究焦点对准韩语文本情感分析。利用语言信息特征, 例如词的语义类别、否定词的修饰范围和词缀的功能意义等。Q. Zhang, Y. Wu 和 T. Li 等^[21]提出将浅层依存分析方法和信息抽取技术相结合的情感句分类模型, 其基本思路是抽取出评价句中的产品特征, 例如性能、产品部件、部件属性、公司名字等信息, 然后通过构建情感词库, 并用它提取情感词特征; 最后通过浅层依存关系树扩展特征, 并训练 SVM 模型实现情感句分类。该方法运用了信息抽取技术改进了模型的特征提取方法, 使得模型有更好的分类效果, 但是该方法比较依赖观点词库。王根和赵军^[22]提出情感句分级模型, 将句子的主客观分类、褒贬分类以及褒贬强度分类统一处理, 如图 3 所示:

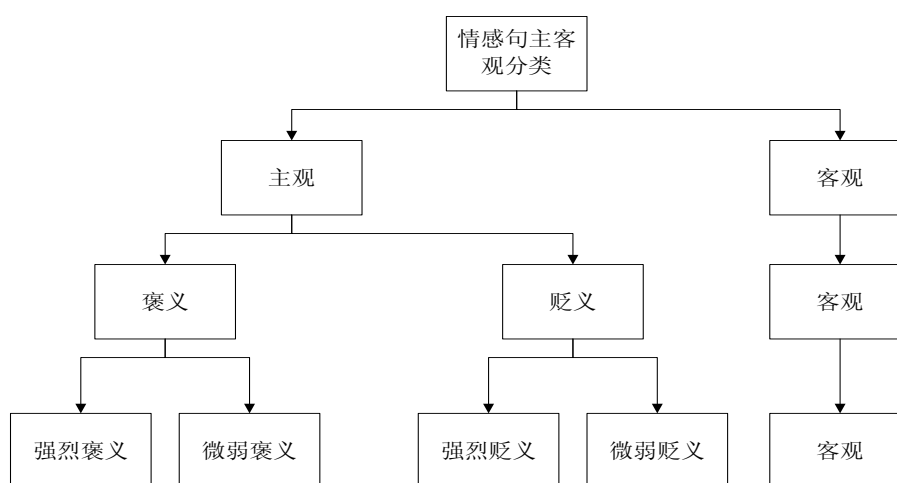


图 3 层级模型

上一级的情感分析结果作为下一级情感分析的输入, 逐层细化情感分析, 但会出现错误从高层向底层传递现象。

2.2.3 基于半监督的句级情感分析模型

R. Socher, J. Pennington 和 E. H. Huang 等^[23]提出基于递归自动编码的半监督情感句分析模型。将句子中词递归的两两合并构建短语树, 提取短语节点特征预测句子情感极性; 该模型不使用任何人工构建的情感词典以及语法分析器等辅助工具, 具有良好的可扩展性, 不局限于二值分类; 但是编码比较复杂, 并且长句编码后节点树比较大、种类比较多, 使得需要计算的节点增加, 分析效率下降。

2.3 篇章级情感分析模型

篇章级情感分析相对于词汇和句子情感分析难度更大。基于无监督的篇章级情感分析与句级情感分析相似, 都是根据词汇之间的极性和强度累加确定情感极性和强度, 从三个方面: 领域适应、特征抽取以及语料不足问题详细介绍基于有监督的篇章级情感分析。基于半监督的篇章级情感分析针对传统模型解决情感分析时存在的问题提出改进模型。

2.3.1 基于无监督的篇章级情感分析模型

M. Taboada, J. Brooke 和 M. Tofiloski 等^[24]提出基于词典的篇章情感分析模型, 即情感极性计算器。该模型使用标注了情感极性和强度的词典, 并且结合副词和否定词分析词汇的情感极性。实验表明在不同领域具有相似的性能。A. L Maas, R. E Daly 和 P. T Pham 等^[25]提出无监督 (获取词汇之间的语义相似信息) 和有监督 (获取词汇的

情感信息)混合模型,利用连续(情感强度不是二值的强弱之分,而是可以量化的连续值)的和多维的情感信息(词汇、短语、句子和篇章)分析篇章情感。不足之处在于假设词汇情感不会受上下文影响。

2.3.2 基于有监督的篇章级情感分析模型

本部分主要从三个角度说明:模型的领域适应问题、特征提取和选择问题以及语料不足问题。在领域适应方面, S. Tan 和 Y. Wang^[26]提出加权的结构对应学习模型。作者重点研究了高频领域专属特征在构建领域分类模型时的不利影响。研究表明高频领域专属特征拥有较大的权重并且承载对应的情感信息。基于共现信息熵获取高频领域专属特征以及构建特征加权模型。实验表明该模型可以提高传统迁移学习模型的分类精度。在 6 个领域的适应性任务中,该模型的性能好于有监督、半监督以及基准的迁移学习模型。尽管该模型可以提高情感分析性能,但是仍存在改进空间。例如:构建高频领域专属特征的最优策略,以及尝试将其他分类器融入这一框架。X. Gloro, A. Bordes 和 Y. Bengio^[27]针对领域适应问题,在原领域使用标注数据训练分类器,并在其他领域扩展该系统。在 22 个领域数据验证了方法的领域适应能力。Q. Wu, S. Tan 和 X. Cheng 等^[28]将研究焦点对准篇章级跨领域情感分析问题,提出借助图排序算法的迭代情感分析模型。构建文本和词汇关系网络,其中关系包括:测试文本和训练文本的关系、测试文本之间的关系、测试文本中词和训练文本中词的关系、测试文本中词之间的关系、测试文本和训练文本中词的关系、测试文本和测试文本中词的关系、测试文本中词和训练文本的关系以及测试文本中词和测试文本的关系;然后采用图排序算法迭代预测各个文本和词的情感极性。该方法利用不同领域的共享信息,但是作者只采用余弦测量法、滑动窗口测量法和向量测量法来计算文本相似,没有考虑句子的结构以及词语之间的相互关系对文本相似的影响。

在特征选择和提取方面, A. Abbasi, S. France 和 Z. Zhang 等^[29]针对整合 n 元特征时噪声、不相关和冗余属性的问题,提出基于规则的多变量特征选择模型,即特征关系网络。该模型考虑了语义信息并且利用了 n 元特征之间的语法关系,可以整合异构的 n 元特征来提高情感分类精度。在 3 个在线评论数据上验证模型,该模型优于现有特征选择方法,它能够选择出重要属性,提高分类精度。A. Agarwal, B. Xie 和 I. Vovsha 等^[30]研究 Twitter 数据的情感分类问题。引入了词性特征,尝试采用树核(Tree Kernel)表示情感特征,以便融入更多的特征。A. Balamurali, A. Joshi 和 P. Bhattacharyya^[31]提出将 WordNet 中词汇的词义作为语义特征,在处理测试数据中未知特征时,采用 WordNet 的相似度计算方法来寻找训练集的特征替代未知特征,然后训练 SVM 分类器预测篇章情感极性,但该方法依赖 WordNet 资源获取特征。S. Li, S. Y. M. Lee 和 Y. Chen 等^[32]构建情感极性转移 SVM 分类器,并用该分类器将语料分成极性转移和极性不转移两个部分,然后训练极性转移和极性不转移 SVM 分类器,最后采用两个分类器联合预测文本。缺点是基准情感词选择易受人为干扰。G. Paltoglou 和 M. Thelwall^[33]在原有的 SVM 分类模型的基础上引入更多复杂的特征以提高分类效果,即引入 tf-idf 特征加权策略。S. Li, R. Xia 和 C. Zong 等^[34]针对特征选择问题,提出基于频率和比例测量方法的加权特征选择策略。该特征选择方法在不同的领域表现良好。S. Tan, Y. Wang 和 X. Cheng 等^[35]提出基于词典和基于学习的情感分析模型;首先在目标领域按照一定比例抽取样本,并根据样本中情感词比例高的极性来标注该文本,然后使用该标注数据在目标领域上训练分类器。该方法采用的抽取样本的策略不是最优策略。A. Abbasi, H. Chen 和 A. Salem^[36]提出基于信息增益的遗传算法完成特征选择,并训练 SVM 模型来预测情感分类。首先利用信息增益抽取部分特征组成特征集合,然后利用遗传算法从该特征集合中选取最有用的特征,最后训练 SVM 模型并预测篇章情感倾向。该方法优点在于将信息增益和遗传算法结合实现有效的选择特征,但是由于遗传算法的随机性,无法保证在规定步数内找出最优特征组合。除此之外,当特征空间巨大时,遗传算法效率较低。R. McDonald, K. Hannan 和 T. Neylon 等^[37]提出了基于结构的从细到粗的篇章情感分析模型。将句子的情感分析结果作为篇章分类器的输入,通过句子情感分析来推理篇章的情感倾向。该方法不局限于两层结构,可以扩展到多层,例如从词汇到句子再到篇章。徐琳宏、林鸿飞和杨志豪^[38]提出基于语义理解的文本情感倾向性分析。首先计算词汇与知网中褒贬词汇间的相似度,获取词汇的情感极性;再选择极性明显的词汇作为特征训练 SVM 分类器;最后基于否定词规则提高分类精度。S. Wang, Y. Wei 和 D. Li 等^[39]提出基于类别显着性词和信息增益的

混合特征提取方法,通过训练 SVM 模型预测情感极性。该方法考虑到领域性强的词对情感极性的影响,但是没有考虑低频名词对篇章极性的影响。

在语料匮乏问题方面, X. Wan^[40]针对缺少中文情感标注资源问题,借助可用的英文资源实现中文情感分类。首先综合基于词典和基于语料的方法优点,使用机器翻译技术构建中文语料,然后提出双语联合训练模型。使用机器翻译技术将标注的英文数据翻译成中文数据,并且将无标注的中文数据翻译成英文数据。然后采用联合训练算法训练中、英文情感分类器。使用两个分类器获得两个情感预测值,最终的情感类别由预测值的平均值决定。但中文的特征抽取精度不准确,且将预测值平均值作为预测最终结果没有给出理论分析。

2.3.3 基于半监督的篇章级情感分析模型

C. Tan, L. Lee 和 J. Tang 等^[41]认为社会关系信息可以提高情感分析精度。假设存在关系的用户可能有相似的观点,提出基于半监督学习和用户关系信息的情感分析模型。用户关系信息包括 Twitter 上的关注关系以及引用关系。实验结果表明该方法性能高于基于 SVM 模型的情感分类,并且对于一个稀疏的关注网络,这种方法仍然有效。存在的问题:(1)该模型没有验证在不同社会网络系统的适应性;(2)没有剔除网络噪声。S. Li, Z. Wang 和 G. Zhou^[42]等针对语料不平衡问题,提出通用的半监督情感分类模型:采用欠采样方法生成均衡的训练样本数据集,然后提出基于随机子空间的半监督学习情感分析模型,但没有确定迭代终止标准。S. Li, C. Huang 和 G. Zhou 等^[43]将研究焦点对准文本情感分类问题,提出了有监督和半监督的分类模型。作者假设个人数据和非个人数据都包含主客观文本,其基本思路:先使用无监督的方法自动将文本 D 分为个人和非个人观点数据,并用这两组观点数据训练两个不同的 SVM 主客观分类器 f_1 和 f_2 以及使用文本 D 训练 SVM 主客观分类器 f_3 ,然后通过固定规则(Fixed Rule)和训练规则(Trained Rule)联合分类器 f_1 、 f_2 和 f_3 ,最后迭代的预测篇章情感极性。

3 情感分析模型应用

情感分析在不同领域具有应用价值,本文将从问答系统、电子商务产品推荐和舆情分析等领域阐述情感分析技术在生活中的相关应用。

3.1 问答系统

M. Mohtarami, H. Amiri 和 M. Lan 等^[44]提出度量答案对问题的不确定性模型。为了推理出答案,借助反义词和同义词消除情感歧义以及确定情感形容词之间的语义关系。方法具体分为四步:(1)度量所有问答对中答案相对于问题的确定性。(2)对于每个问答对,通过计算得到一个阈值,该阈值评价答案对于问题的确定性。(3)根据答案的确定性和阈值推理答案。(4)使用同义词细化方法。对于步骤(1),根据答案和问题中的形容词之间的共现关系计算确定性:关系强度越大,确定性越大。对于步骤(2),因为反义词与该词属于相同的情感强度等级,所以假设答案中的形容词和问题中的形容词的关系强度大于答案中的形容词和问题中的形容词的反义词的关系强度,则答案对于问题相对确定,即答案中的形容词与问题中的形容词关系更强,则这个答案更趋向于肯定回答。反义词根据资源获得,例如 WordNet。对于步骤(3)根据阈值判断答案是(或否):大于阈值则“是”,小于为“否”,其他为不确定。对于步骤(4),使用同义词工具扩展答案中的形容词集合。在度量确定性时,通过度量扩展后的形容词集合与问题中形容词的强度,以及与问题中形容词的反义词的强度最终确定答案。

3.2 产品推荐

Bing Qin, Yanyan Zhao 和 Leilei Gao 等^[45]将研究焦点对准商品个性化推荐问题,首先通过基于情感词提取特征,通过句法分析来获取句法特征,最后训练最大熵模型预测评价语句的情感极性。观点判断主要分为两个步骤:(1)首先通过情感分类器获取句子的情感极性以及强度,根据强度阈值选择情感语句作为该商品的推荐原因;(2)对于每一个产品,统计其分析结果中褒义评价所占比例,并将比例超过 50%的产品推荐给用户。

3.3 舆情分析

M. Thelwall, K. Buckley 和 G. Paltoglou^[46]研究一个月的 Twitter 数据, 实验表明公众事件可以强化情感。一方面, 通过对 30 个流行事件研究, 结果显示公众事件可以强化人们的负向情感。另一方面, 最让人感兴趣的事件会增加人们的积极情感。

3.4 其他

通过情感分析技术, 对网络上有害身心健康的、泄露国家机密的、侮辱国家名誉的以及传播淫秽色情的信息进行监控拦截, 维护网络的安全和营造一个和谐干净的网络世界。通过提取用户对商品的评价信息可以帮助公司了解目前哪些商品是受用户好评的, 以及商品的优缺点。网络存在很多网络社区, 例如 QQ、人人、微薄、以及一些 BBS 等, 网友们无时无刻在上面发表言论。通过融入情感分析技术, 可以发掘他们的喜好, 挖掘当中的商机以服务社会。

4 结论

现有的心理学情感分析方法主要从定性角度分析情感, 没有对情感进行量化分析。本文对情感分析量化模型的研究现状与进展进行了综述, 其中重点介绍了情感分析模型中三种关键模型, 包括词级情感分析模型、句级情感分析模型和篇章级情感分析模型。模型的相关应用重点从问答系统、产品推荐和舆情分析三个方面详细介绍。情感分析是语义分析问题, 而解决语义分析问题的前提是语句的深层次结构分析, 因而解决情感分析问题要从语句的深层次结构分析入手。此外, 情感分析属于应用心理学范畴, 但目前情感分析模型没有心理学理论作为理论基础, 我们认为未来构建的模型应以心理学理论为基础。本文综述的模型采用自动化方法定量的分析文本中的情感, 推动了应用心理学发展, 将心理学从定性分析方法引入到定量分析方法, 可以自动量化情感这一不易把握和难以研究的对象。

参考文献

- [1] 赵坤, 张林. 心理学导论. 中国传媒大学出版社. 2009.
- [2] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 271-278.
- [3] S. Kim and E. Hovy, "Determining the Sentiment of Opinions," In Proceedings of COLING, 2004, pp. 1367-1373.
- [4] Y. Mao and G. Lebanon, "Isotonic Conditional Random Fields and Local Sentiment Flow," In Proceedings of NIPS, 2006. pp: 961-968.
- [5] 宋晓雷, 王素格, 李红霞等. 基于概率潜在语义分析的词汇情感倾向判别. 中文信息学报. 2011.
- [6] Y. Wu and M. Wen, "Disambiguating Dynamic Sentiment Ambiguous Adjectives," Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Aug. 2010, pp. 1191-1199.
- [7] A. Hassan and D. Radev, "Identifying Text Polarity Using Random Walks," Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Jul. 2010, pp. 395-403.
- [8] 董喜双, 邹启波, 关毅等. 基于最大熵模型和最小割模型的中文词与句褒贬极性分析. 第三届中文倾向性分析评测. 2011.
- [9] D. Das and S. Bandyopadhyay, "Word to Sentence Level Emotion Tagging for Bengali Blogs," Proceeding ACL Short '09 Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Aug, 2009, pp. 149-152.
- [10] D. Rao and D. Ravichandran, "Semi-Supervised Polarity Lexicon Induction," Proceedings of the 12th Conference of the European Chapter of the ACL, 2009, pp. 675-682.
- [11] Blum, Lafferty, Rwebangira, et al. "Semi-supervised Learning Using Randomized Mincuts," In Proceedings of the ICML. 2004.
- [12] X. Zhu and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation," Technical Report CMU-CALD-02-107, Carnegie Mellon University. 2002.
- [13] G. Fu and X. Wang, "Chinese Sentence-level Sentiment Classification based on Fuzzy Sets," Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Aug, 2010, pp. 312-319.
- [14] A. Meena, T. Prabhakar, G. Amati, et al. "Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis," Advances in Information Retrieval, vol. 4425. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 573-580.
- [15] H. Guo, H. Zhu, Z. Guo, and Z. Su, "Domain Customization for Aspect-oriented Opinion Analysis with Multi-level Latent Sentiment

- Clues,” in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 2493-2496.
- [16] R. Xia and C. Zong, “Exploring the Use of Word Relation Features for Sentiment Classification,” Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Aug, 2010, pp. 1336-1344.
- [17] B. Wei and C. Pal, “Cross Lingual Adaptation: An Experiment on Sentiment Classifications,” Proceeding ACL Short '10 Proceedings of the ACL 2010 Conference Short Papers, Jul, 2010, pp. 258-262.
- [18] D. Davidov, O. Tsur, and A. Rappoport. “Enhanced Sentiment Learning Using Twitter Hashtags and Smileys,” Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp.241-249.
- [19] W. Wei and J. A. Gulla, “Sentiment Learning on Product Reviews via Sentiment Ontology Tree,” Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Jul, 2010, pp. 404-413.
- [20] H. Jang and H. Shin, “Language-specific Sentiment Analysis in Morphologically Rich Languages,” Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Aug, 2010, pp. 498-506.
- [21] Q. Zhang, Y. Wu, T. Li, et al, “Mining Product Reviews based on Shallow Dependency Parsing,” in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 726-727.
- [22] 王根,赵军. 基于多重标记 CRF 句子情感分析的研究.全国第九届计算语言学学术会议. 2007.
- [23] R. Socher, J. Pennington, E. H. Huang, et al, “Semi-supervised Recursive Auto-encoders for Predicting Sentiment Distributions,” Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Jul, 2011, pp. 151-161.
- [24] M. Taboada, J. Brooke, M. Tofiloski, et al, “Lexicon-Based Methods for Sentiment Analysis,” Computational Linguistics, vol. 37, no. 2, Jun, 2011, pp. 267-307.
- [25] A. L Maas, R. E Daly, P. T Pham, et al, “Learning Word Vectors for Sentiment Analysis,” Computational Linguistics, vol. 31, no. Jne, 2011, pp. 142-150.
- [26] S. Tan and Y. Wang, “Weighted SCL Model for Adaptation of Sentiment Classification,” Expert Systems with Applications, vol. 38, no. 8, Aug, 2011, pp. 10524-10531.
- [27] X. Glorot, A. Bordes, and Y. Bengio, “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach,” Proceedings of the 28th International Conference on Machine Learning, 2011, pp.513-520.
- [28] Q. Wu, S. Tan, X. Cheng, et al, “MIEA: a Mutual Iterative Enhancement Approach for Cross-Domain Sentiment Classification,” Coling 2010 Posters, no. Aug, 2010, pp. 1327-1335.
- [29] A. Abbasi, S. France, Z. Zhang, et al, “Selecting Attributes for Sentiment Classification Using Feature Relation Networks,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 3, Mar, 2011, pp. 447-462.
- [30] A. Agarwal, B. Xie, I. Vovsha, et al, “Sentiment Analysis of Twitter Data,” Proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media, Jun, 2011, pp. 30-38.
- [31] A. Balamurali, A. Joshi, and P. Bhattacharyya, “Harnessing WordNet Senses for Supervised Sentiment Classification,” EMNLPACL, 2011, pp. 1081-1091.
- [32] S. Li, S. Y. M. Lee, Y. Chen, et al, “Sentiment Classification and Polarity Shifting,” Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Aug, 2010, pp. 635-643.
- [33] G. Paltoglou and M. Thelwall, “A Study of Information Retrieval Weighting Schemes for Sentiment Analysis,” Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Jul, 2010, pp. 1386-1395.
- [34] S. Li, R. Xia, C. Zong, et al, “A Framework of Feature Selection Methods for Text Categorization,” Proceeding ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Aug, 2009, pp. 692-700.
- [35] S. Tan, Y. Wang, and X. Cheng, “Combining Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labeled Examples,” in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 743-744.
- [36] A. Abbasi, H. Chen, and A. Salem, “Sentiment Analysis in Multiple Languages,” ACM Transactions on Information Systems, vol. 26, no. 3, Jun, 2008, pp.1-34.
- [37] R. Mcdonald, K. Hannan, T. Neylon, et al, “Structured Models for Fine-to-Coarse Sentiment Analysis,” Proceedings of the Association for

Computational Linguistics, 2007, pp. 432-439.

- [38] 徐琳宏,林鸿飞,杨志豪.基于语义理解的文本倾向性识别机制.中文信息学报.2007.
- [39] S. Wang, Y. Wei, D. Li, et al, "A Hybrid Method of Feature Selection for Chinese Text Sentiment Classification," in Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 2007, pp. 435-439.
- [40] X. Wan, "Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews," Computational Linguistics, vol. 37, no. 3, Sep, 2011, pp. 587-616.
- [41] C. Tan, L. Lee, J. Tang, et al, "User-level Sentiment Analysis Incorporating Social Networks," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 1397-1405.
- [42] S. Li, Z. Wang, G. Zhou, et al, "Semi-Supervised Learning for Imbalanced Sentiment Classification", in Proc. IJCAI, 2011, pp.1826-1831.
- [43] S. Li, C.-R. Huang, G. Zhou, et al, "Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification," Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Jul, 2010 pp. 414-423.
- [44] M. Mohtarami, H. Amiri, M. Lan, et al, "Predicting the Uncertainty of Sentiment Adjectives in Indirect Answers," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 2485-2488.
- [45] B. Qin, Y. Zhao, L. Gao, et al, "Recommended or Not? Give advice on Online Product," Fifth international conference on Fuzzy Systems and Knowledge Discovery, 2008, pp.208-212.
- [46] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter Events," Journal of the American Society for Information Science and Technology, vol. 62, no. 2, Feb. 2011, pp. 406-418.

【作者简介】



董喜双（1981-），男，汉族，博士生，自然语言处理。2001 年至 2005 年在哈尔滨理工大学计算机科学与技术系学习，2005 年获工学学位，2005 年至 2008 年在哈尔滨工程大学计算机软件与理论专业学习，期间作为主要研制人员参加“海洋工程公司信息管理系统”的开发工作，2008 年至今年在哈尔滨工业大学计算机应用专业攻读博士学位，作为第一负责人参加“博客或 BBS 情感倾向性分析系统”研发工作（与富士通研究开发中心有限公司合作），dongxishuang@gmail.com。



邹启波（1993-），男，汉族，本科生，自然语言处理。2009 年至今在哈尔滨工业大学软件工程系学习，2010 参加开发哈工大软件学院成绩分析和追踪系统，2010 年被评为优秀团员以及荣获人民奖学金，2011 年荣获黑龙江省大学生程序设计竞赛(ACM/ICPC)二等奖，2011 年参加第三届中文倾向性分析评测(COAE2011)，其中词级任务准确率达评测最高值。2011 年发表论文《基于最大熵模型和最小割模型的中文词与句褒贬极性分析》，zouqibo2009@163.com。



关毅（1970-），男，汉族，博士，教授，自然语言处理。1988 年至 1992 年在天津大学计算机科学与工程系软件专业学习，1992 年 6 月获工学学士学位。1992 年至 1995 年在哈尔滨工业大学计算机应用专业攻读硕士学位，1995 年获得硕博连读资格。1995 年至 1999 年在哈尔滨工业大学计算机应用专业攻读博士学位，其间作为主要研制人之一参加了“微软拼音输入法”（与微软公司合作）和“BOPOMOFO”智能汉字输入系统（与日本佳能公司合作）的研制工作，1996 年 3 月至 1997 年 3 月赴日本佳能泰克公司研修一年，1999 年 3 月在哈尔滨工业大学计算机学院获得计算机应用专业工学博士学位，guanyi@hit.edu.cn。